ISSN(Online) : 2456-8910
International Journal of Innovative Research in Applied Sciences and Engineering (IJIRASE)
Volume 4, Issue 2, DOI: 10.29027/IJIRASE.v4.i2.2020.629-632, August 2020

# Used Car Price Prediction using K-Nearest Neighbor Based Model

K.Samruddhi[1], Dr R.Ashok Kumar[2]
*Department Of ISE*
*B.M.S College of Engineering,*
*Affiliated to VTU*
Bangalore, Karnataka
Email: samruddhi96.sk@gmail.com, ashokkumar.ise@bmsce.ac.in

***Abstract*—** Predicting the price of used cars is one of the significant and interesting areas of analysis. As an increased demand in the second-hand car market, the business for both buyers and sellers has increased. For reliable and accurate prediction it requires expert knowledge about the field because of the price of the cars dependent on many important factors. This paper proposed a supervised machine learning model using KNN (K Nearest Neighbor) regression algorithm to analyze the price of used cars. We trained our model with data of used cars which is collected from the Kaggle website. Through this experiment, the data was examined with different trained and test ratios. As a result, the accuracy of the proposed model is around 85% and is fitted as the optimized model.

***Keywords— K Nearest Neighbor, Prediction, Machine Learning, Used Cars Accuracy, Preprocessing, Regression, Cross-validation, K-Fold.***

## I. INTRODUCTION

The second-hand car market has continued to expand even as the reduction in the market of new cars. According to the recent report on India's pre-owned car market by Indian Blue Book, nearly 4 million used cars were purchased and sold in 2018-19. The second-hand car market has created the business for both buyers and sellers. Most of the people prefer to buy the used cars because of the affordable price and they can resell that again after some years of usage which may get some profit. The price of used cars depends on many factors like fuel type, colour, model, mileage, transmission, engine, number of seats etc., The used cars price in the market will keep on changing. Thus the evaluation model to predict the price of the used cars is required.

In this paper, we proposed a model to estimate the cost of the used cars using the K nearest neighbour algorithm which is simple and suitable for small data set. Here, we have collected a used cars dataset and analyzed the same. The data was trained by the model and we examined the accuracy of the model among different ratios of trained and test set. The same model is cross-validated for assessing the performance of the model using the K- Fold method which is easy to understand and implement.

The paper structured in the following manner: Section II contains the literature survey related to the field of used cars price prediction. In section III the methodology of the study was proposed. Section IV elaborates the examination of the performance of the model and cross-validation of the proposed model for price prediction of the used cars. Finally, Section V specifies the conclusion and future work.

## II. LITERATURE SURVEY

Sameerchand Pudaruth[1] proposed Predicting the Price of Used Cars using Machine Learning Techniques. In this paper, they collected the historical data of used cars in Mauritius from the newspapers and applied different machine learning techniques like decision tree, K-nearest neighbours, Multiple Linear Regression and Naïve Bayes algorithms to predict the price. This model has the mean error about Rs27000 for Nissan cars and about Rs45000 for Toyota cars using KNN and around Rs51000 using linear regression. The accuracy of decision trees and Naïve Bayes algorithm dangled between 60 to 70 percentile with different parameters and the overall training accuracy of the model is 61%.

Nitis Monburinon et al. [2] proposed prediction of Prices for Used Car by Using Regression Models. In this paper, the authors selected the data from the German e-commerce site. The main goal of this work is to find a suitable predictive model to predict the used cars price. They used different machine learning techniques for comparison and used the mean absolute error(MAE) as the metric. They proposed that their model with gradient boosted regression has a lower error with MAE value 0.28 and this gives the higher performance where linear regression has the MAE value 0.55, random forest with MAE value 0.35.

Enis Gegic et al. [3] proposed Car Price Prediction using Machine Learning Techniques. In this paper, they proposed an ensemble model by collecting different types of machine learning techniques like Support Vector Machine, Random Forest and Artificial neural network. They collected the data from the web portal www.autopijaca.ba and build this model to predict the price of used cars in Herzegovina and Bosnia. The accuracy of their model is 87%.

Kanwal Noor and Sadaqat Jan[4] proposed Vehicle Price Prediction System using Machine Learning Techniques. In this paper, they proposed a model to predict the price of the cars through multiple linear regression method. They selected the most influencing feature and removed the rest by performing feature selection technique. The Proposed model achieved the prediction precision of about 98%.

In this paper, a machine learning model is proposed to estimate the cost of the used cars using K-Nearest Neighbor algorithm. The model is trained with used cars data for different trained and test ratios. Then the proposed model is cross-validated using K fold method to examine the performance to avoid the overfit.

## III. METHODOLOGY

The Used Cars data set was taken and data processing has done to filter the data and to remove some unnecessary data. The model was trained with the processed data using KNN algorithm to predict the sales of used cars with higher accuracy. Fig 1 shows the structured outline for proposed Methodology.

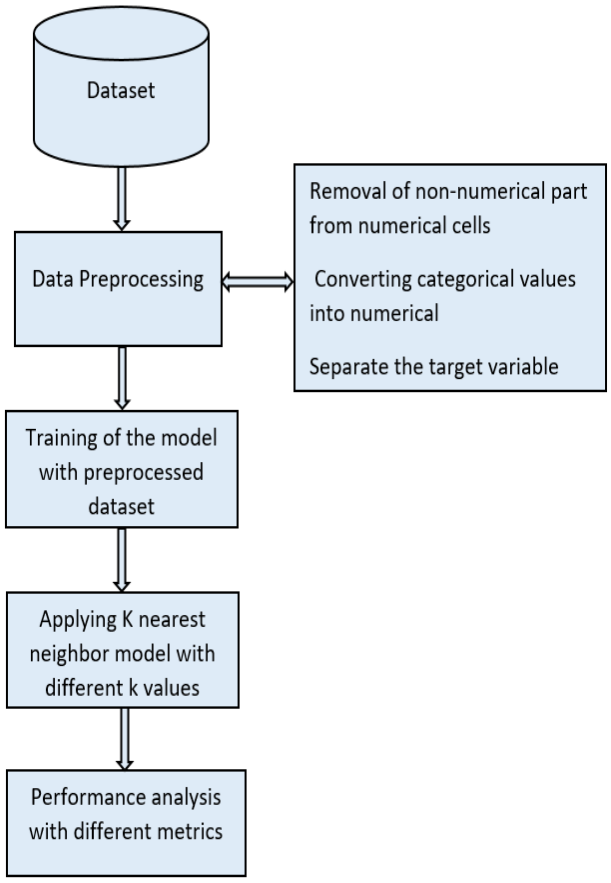


Fig 1: Structured outline of Proposed Methodology

### *A) Dataset Collection*

It is the process of gathering the information from the source for the evaluation. The Used Cars data set is collected from a website Kaggle which is in a CSV format. The data set contains 14 variables which include an unnamed serial number, Name,Location,ileage, Fuel_Type, Engine Transmission, Kilometers_Driven, Power, New_Price,Year, Seats, Owner_Type, Price as shown in Fig 2.

### *B) Data Preprocessing*

This step is one of the important steps in supervised machine learning. It includes the following.

*i) Removal of Non-numerical part from numerical features*

This step removes the non-numerical words from the features like Mileage, Engine, and power for data processing.

*Step1: Converting the data frame into a list.*

*Step2: Splits the list based on a delimiter.*

*Step3: Store the required data back to the data frame.*

*ii) Converting Categorial values into numerical*

Here, the categorical values like Name, Location, Fuel_Type, Transmission, Owner_Type are converted to numerical because machine learning deal with numerical values easily because of the machine-readable form. This is done by using Label Encoder which is a python package.

*Step1: We have to select categorical values based on its datatype.*

*Step2: Converting the categorical values into numerical values by using Label encoder API in python.*

*iii) Separate the target variable*

Here, we have to separate the target feature which is we are going to predict. In this case, price is the target variable.

*Step1: The target variable price is assigned to the variable 'y'.*

*Step2: The preprocessed data set except the target variable is assigned to the variable 'X'.*

After all preprocessing steps have done, the data was shown as in figure Fig 3.

### *c) K Nearest neighbour (KNN)*

In this work, we have used K Nearest Neighbor algorithm to prepare a model which predict the price of the used cars. By using KNN, it is easy to implement machine learning models. It is a non-parametric method used for both regression and classification. It estimates the numerical target based on a similarity measure. A simple implementation of KNN is to find the average of the numerical target of the K nearest neighbours.

| | Unnamed: 0 | Name | Location | Year | Kilometers_Driven | Fuel_Type | Transmission | Owner_Type | Mileage | Engine | Power | Seats | New_Price | Price |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | Maruti Wagon R LXI CNG | Mumbai | 2010 | 72000 | CNG | Manual | First | 26.6 km/kg | 998 CC | 58.16 bhp | 5.0 | NaN | 1.75 |
| 1 | 1 | Hyundai Creta 1.6 CRDi SX Option | Pune | 2015 | 41000 | Diesel | Manual | First | 19.67 kmpl | 1582 CC | 126.2 bhp | 5.0 | NaN | 12.50 |
| 2 | 2 | Honda Jazz V | Chennai | 2011 | 46000 | Petrol | Manual | First | 18.2 kmpl | 1199 CC | 88.7 bhp | 5.0 | 8.61 Lakh | 4.50 |
| 3 | 3 | Maruti Ertiga VDI | Chennai | 2012 | 87000 | Diesel | Manual | First | 20.77 kmpl | 1248 CC | 88.76 bhp | 7.0 | NaN | 6.00 |
| 4 | 4 | Audi A4 New 2.0 TDI Multitronic | Coimbatore | 2013 | 40670 | Diesel | Automatic | Second | 15.2 kmpl | 1968 CC | 140.8 bhp | 5.0 | NaN | 17.74 |

Fig 2: Sample data before data preprocessing

| | Year | Kilometers_Driven | Mileage | Engine | Power | Seats | Price | Name | Location | Fuel_Type | Transmission | Owner_Type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2010 | 72000.0 | 26.60 | 998.0 | 58.16 | 5.0 | 1.75 | 18 | 9 | 0 | 1 | 0 |
| 1 | 2015 | 41000.0 | 19.67 | 1582.0 | 126.20 | 5.0 | 12.50 | 10 | 10 | 1 | 1 | 0 |
| 2 | 2011 | 46000.0 | 18.20 | 1199.0 | 88.70 | 5.0 | 4.50 | 9 | 2 | 3 | 1 | 0 |
| 3 | 2012 | 87000.0 | 20.77 | 1248.0 | 88.76 | 7.0 | 6.00 | 18 | 2 | 1 | 1 | 0 |
| 4 | 2013 | 40670.0 | 15.20 | 1968.0 | 140.80 | 5.0 | 17.74 | 1 | 3 | 1 | 0 | 2 |

Fig 3: Sample data after data preprocessing

Here, we have estimated the accuracy of the model by training with different values of k from 2 to 10 to find a comparative as best performance. Here prediction is made by looking whole training set to find the k most similar values. To find the most similar values for the new data, the distance is measured using the Euclidean Distance metric. Then the average of the measure is taken to find the estimation value. The formula for Euclidean Distance is as shown below where A and B are two points for which the distance should be calculated.

$$dist(A,B) = \sqrt{\frac{\sum_{i=1}^{m}(x_i - y_i)^2}{m}}$$

## IV. Experimentation Results

In this work, the model is trained with data for 3 different ratios for k values from 2 to 10. It is getting 85% accuracy, Root-Mean Squared Error (RMSE) rate of 4.01 and Means Absolute Error (MAE) rate of 2.01 with K value of 4. The accuracy of the model with different k values of KNN with different data ratios is shown in Table 1 and the same is plotted in the figure Fig4.

Table 1: Accuracy for different trained and test ratios for k values from 2 to 10

| K value for KNN | Accuracy for different ratios | | |
|---|---|---|---|
| | Training =75% & Testing = 25% | Training =80% & Testing = 20% | Training =85% & Testing = 15% |
| K=2 | 0.809 | 0.844 | 0.842 |
| K=3 | 0.813 | 0.848 | 0.848 |
| K=4 | 0.802 | 0.845 | **0.85** |
| K=5 | 0.798 | 0.834 | 0.833 |
| K=6 | 0.703 | 0.837 | 0.817 |
| K=7 | 0.795 | 0.826 | 0.807 |
| K=8 | 0.788 | 0.819 | 0.798 |
| K=9 | 0.785 | 0.809 | 0.788 |
| K=10 | 0.788 | 0.806 | 0.785 |

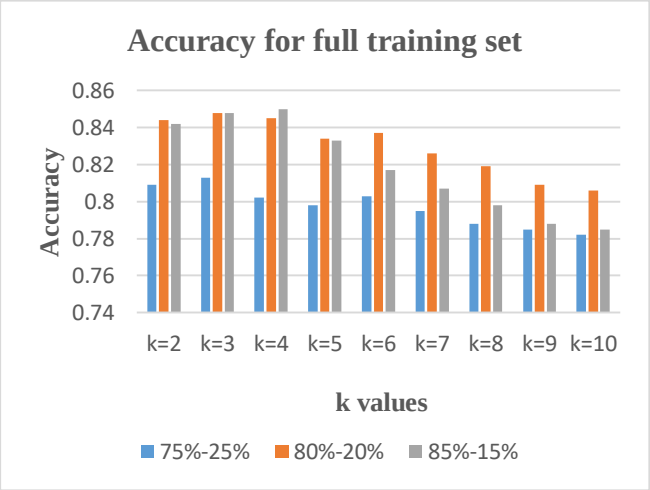


Fig 4: Accuracy for the full training data with different trained and test ratios

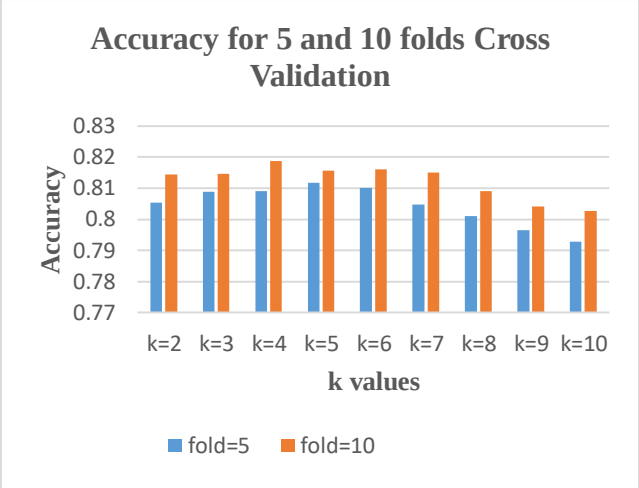


Fig 5: Accuracy for 5 and 10 folds cross-validation with different k values of KNN

**Cross-validation:** When we randomly select the training and test set, it may give high accuracy and when we select another train and test data set it may give lower accuracy. So Cross-validation is used to inspect the overfitting of the model. It helps us to know about how our predictive model can show the result for the new dataset.

In this work, k fold cross-validation is evaluated for the predictive model. In this, it splits the data set into k subparts and take one part as test data then the remaining part as the trained data. Then it will repeat the same steps for n times and produce n results. The average of the result has taken for the estimation value.

Here, we have validated our proposed model for 5 folds and 10 folds. It is getting accuracy 82%, RMSE rate 4.73 and MAE rate 2.13 for 10 folds with K value of 4. It is seen that the proposed model getting the best results after cross-validation. Fig 5 shows the accuracy for different k values of KNN with 5 and 10 folds as listed in Table 2.

Table 2: Accuracy for 5 and 10 folds for k values from 2 to 10

| K value for KNN | Accuracy for 5 folds and 10 folds Cross-validation | |
|---|---|---|
| | Fold = 5 | Fold = 10 |
| K=2 | 0.805 | 0.814 |
| K=3 | 0.808 | 0.814 |
| K=4 | 0.809 | **0.82** |
| K=5 | 0.811 | 0.815 |
| K=6 | 0.81 | 0.816 |
| K=7 | 0.804 | 0.815 |
| K=8 | 0.801 | 0.809 |
| K=9 | 0.796 | 0.804 |
| K=10 | 0.792 | 0.802 |

## V. Conclusion

In this paper, we have trained our model with used cars data set to predict the price. Here we have used K nearest Neighbor algorithm and we got accuracy 85% where the accuracy of linear regression is 71%. The proposed model is also validated with 5 and 10 folds by using K Fold Method. The experimental analysis shows that the proposed model is fitted as the optimized model.

In our future work, we will apply advanced machine learning techniques and validate the model with different methods to enhance the optimization of the model with improved accuracy.

## VI. References

[1] Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." *Int. J. Inf. Comput. Technol* 4, no. 7 (2014): 753-764.

[2] Monburinon, Nitis, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, and Pitchayakit Boonpou. "Prediction of prices for used car by using regression models." In *2018 5th International Conference on Business and Industrial Research (ICBIR)*, pp. 115-119. IEEE, 2018.

[3] Gegic, Enis, Becir Isakovic, Dino Keco, Zerina Masetic, and Jasmin Kevric. "Car price prediction using machine learning techniques." *TEM Journal* 8, no. 1 (2019): 113.

[4] Noor, Kanwal, and Sadaqat Jan. "Vehicle price prediction system using machine learning techniques." *International Journal of Computer Applications* 167, no. 9 (2017): 27-31.

[5] https://www.kaggle.com/datasets

[6] https://ieeexplore.ieee.org/Xplore/home.jsp

[7]https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

[8]https://machinelearningmastery.com/k-fold-cross-validation/

[9]https://kite.com/python/docs/sklearn